IE 529
Homework set 3: due Friday, October 27

1. As we have and will encounter Jensen's inequality and the *geometric-mean algebraic-mean* (GM-AM) inequality in our readings, we will work through the details of these in this homework problem.

   Preliminary: A subset D of a real vector space (e.g., $\mathbf{R}^d$) is convex (concave) if every convex (concave) linear combination of a pair of points of D is in D, i.e., if $x, y \in D$ and $0 < \alpha < 1$ imply that $\alpha x + (1-\alpha)y \in D$. A function $f : D \to \mathbf{R}$ is similarly said to be convex (concave) if $f(\alpha x + (1-\alpha)y) \leq (\geq) \alpha f(x) + (1-\alpha) f(y)$. These notions can be extended to linear combinations of any finite number of points, with scalings α_i such that $\sum_i \alpha_i = 1$.

   Prove the following.
   Jensen's inequality: Suppose the function $f : D \to \mathbf{R}$ is a concave function. Assume $x_1, x_2, \ldots, x_n \in D$ and $0 < \alpha_i < 1$ for $i = 1, 2, \ldots, n$ with $\sum_i \alpha_i = 1$. Then

   $$\sum_{i=1}^{n} \alpha_i f(x_i) \leq f\left(\sum_{i=1}^{n} \alpha_i x_i\right).$$

   Hints: First note for the case $n = 1$ there is nothing to prove and for $n = 2$ the statement follows immediately from the definitions. So consider $n \geq 3$ and an induction argument. That is, assume the statement is true for some small n, and show it holds for $n + 1$.

   When will equality hold?

2. Now using Jensen's show the **GM-AM inequality** holds:
   Let $\{x_i\}$, $i = 1, 2, \ldots n$, be a set of n non-negative real numbers. Show that the following inequality holds:

   $$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \leq \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

   Hint: note that the function $f(x) = \log x$ is concave on $(0, \infty)$.

3. (Prob. 29 in Ross text) The regression model $Y = \beta x + e$, for $e \in N(0, \sigma^2)$, is called regression through the origin, as it presupposes that the expected response corresponding to the input level $x = 0$ is 0.
   Suppose that (x_i, Y_i), $i = 1, \ldots, n$, is a data set from this model.

   (a) Determine the least squares estimator $\hat{\beta}$ of β.

   (b) What is the distribution of $\hat{\beta}$?

   (c) Write an expression for the resulting sum-of-square-error criterion.

   (d) Construct a hypothesis test framework for: $H_0 : \beta = \beta_0$ versus $H_a : \beta \neq \beta_0$.

4. (Prob. 46 in Ross text) The following data resulted following a series of Stanford heart transplants. This data relates *survival time* (in days) of heart transplant recipients, to their *age at time of transplant*, and to a so-called *mismatch score* that supposedly indicates fit of donor and recipient.

Survival time	Mismatch score	Age
624	1.32	51.0
46	.61	42.5
64	1.89	54.6
1,350	.87	54.1
280	1.12	49.5
10	2.76	55.3
1,024	1.13	43.4
39	1.38	42.8
730	.96	58.4
136	1.62	52.0
836	1.58	45.0
60	.60	64.5

   (a) Let the dependent variable be the logarithm of Survival time. Fit a multiple linear regression on the independent variables of Mismatch score and Age.

   (b) Compute an estimate of the variance of the error term.

5. (Prob. 58 in Ross text) Twelve first-time heart attack patients were given a test that measures "internal anger". The following data relates their scores, and whether they had a second heart attack within 5 years.

Anger Score	Second Heart Attack
80	Yes
77	Yes
70	No
68	Yes
64	No
60	Yes
50	Yes
46	No
40	Yes
35	No
30	No
25	Yes

(a) Explain how the relationship between a second heart attack and one's anger score can be analyzed via a logistic regression model.

(b) Using a software package of your choice, estimate parameters for this model (for example, in Matlab to fit a logistic model consider the command 'glmfit').

(c) Estimate the probability that a heart attack patient with an anger score of 55 will have a second heart attack within 5 years.

6. On the course website you will find a data file called PCAdata.mat (Matlab format), or PCAdata.csv (Python format).

(a) For this data set compute the SVD (singular value decomposition) of the original matrix, and using this SVD discuss the expected results of performing a PCA on this data.

(b) Compute the PCA: First compute the mean(s) for the data, and subtract from the original data; second compute the covariance matrix including the scaling $1/(n-1)$; third compute an eigenvalue decomposition and sort both the eigenvalues and eigenvectors in descending order.

(c) Plot and discuss the principal components. Discuss how this process and results might differ from a direct SVD of the de-biased, scaled data.